期刊论文核心研究主题识别及其演化路径可视化方法研究*
——以我国医疗健康信息领域期刊论文为例

■ 岳丽欣 周晓英 陈旖旎
中国人民大学信息资源管理学院 北京 100872

摘 要：[目的/意义]提出领域核心研究主题识别及其演化路径可视化方法，以期为领域主题演化分析研究提供借鉴，对于揭示领域核心主题的演化特征与发展规律具有一定的意义。[方法/过程]利用 LDA 模型进行主题识别，结合多维尺度分析和可视化技术将 LDA 主题识别结果映射到二维空间，识别主题之间的关联关系，确定核心主题、次要主题；利用主题相似度算法探测相邻时期主题之间的关联，提出一种新的可视化展示方法，构建不同类型研究主题的交叉演化路径，以揭示核心主题、次要主题在演化过程中的动态变化。[结果/结论]以我国医疗健康信息领域为例进行实证研究，研究结果发现，我国医疗健康信息领域核心研究主题主要有电子健康档案、互联网医疗等，其中，健康管理、智慧医疗等核心主题呈现良好的发展演化趋势。

关键词：核心研究主题 主题识别方法 主题演化路径 可视化方法 医疗健康信息
分类号：G251.2
DOI：10.13266/j.issn.0252-3116.2020.05.010

近年来，利用关键词词频分析(Keyword frequency analysis)、共词分析(Co-word analysis)、引文分析(Citation analysis)、主题探测与追踪(Topic detection and tracking)和主题演化(Topic evolution)等不同的方法来分析领域中的研究主题及其发展趋势[1]，成为国内外情报学研究的热点。但现有的研究多基于 Citespace、Ucinet 和 SPSS 等软件工具进行学术论文的研究主题识别及其发展趋势分析，随着数据量以及用户细粒度需求的变化，共词网络、引文网络等宏观、静态结果逐渐难以满足学科情报分析需求[2]。

目前，基于主题类型划分的研究主题演化动态路径及其时序变迁的研究成果较少，为弥补这一不足，本文首先研究了核心研究主题识别及其演化路径可视化分析方法，提出了构建核心主题和次要主题两种不同类型研究主题的交叉演化路径的一种新的可视化方法，以CNKI 期刊全文数据库医疗健康信息领域的论文数据进行实证分析，并应用可视化分析结果具体分析了医疗健康信息领域核心主题和次要主题的演变过程。

1 文献综述

1.1 主题识别

主题识别是指利用文献计量、自然语言处理等方法技术对科技文献中的研究主题进行挖掘分析。目前主题识别主要有基于共词网络、社区探测和 LDA(Latent Dirichlet Allocation)主题模型等几种方法和模型，相关研究如：A. D. Ritzhaupt 等[3]运用共词网络分析方法进行主题识别，并利用该方法分析了北美地区远程教育的主要研究主题及其发展趋势；程齐凯等[4]提出基于社区探测模型的主题识别方法；王效岳等[5]提出基于 LDA 模型的学科主题识别方法，并利用美国国家自然基金资助的基金项目数据进行了实证研究，验证了方法的可行性。

1.2 主题演化

主题演化是指期刊论文中蕴含的研究主题在时间维度上的动态变化过程，它主要描述了某领域研究主题在一定时间内的成长、分裂、融合、衰退等状态，有助

* 本文是国家自然科学基金项目“医疗健康网站信息可信度与质量控制研究”(项目编号:71473260)和国家社会科学基金项目“健康中国建设中的国民健康促进和健康服务策略研究”(项目编号:16AZD021)研究成果之一。
作者简介：岳丽欣(ORCID:0000-0002-7268-7871)，博士研究生；周晓英(ORCID:0000-0002-9116-1525)，教授，博士生导师，通讯作者，E-mail:xyz-ruc@qq.com；陈旖旎(ORCID:0000-0002-1487-8186)，博士研究生。
收稿日期:2019-05-09 修回日期:2019-10-21 本文起止页码:89-99 本文责任编辑:杜杏叶

于揭示研究的现状、变化和趋势。如何从海量的学术论文中准确、有效地识别研究主题的演化脉络并进行可视化展示成为目前亟需解决的问题。目前很多学者开展了主题演化研究,研究成果如:李湘东等[6]提出一种基于LDA模型的科技期刊主题演化分析方法,引入时间因素,基于LDA主题识别及JS散度计算结果实现主题在强度、内容两方面的演化;刘自强等[7]提出了多维度视角下的主题演化分析方法,构建了主题强度、主题结构和主题内容三个维度的主题演化模型,通过对国内图情领域的大数据研究领域的实证研究验证了该方法的准确性和有效性;周源等[8]将期刊论文外部特征(作者)融入主题分析中,基于加权雅可比相似度算法构建了作者-主题的演化模型,能够分析某一研究主题在不同时期下的影响力较高的学者。

1.3 主题演化路径可视化

数据挖掘、可视化领域的研究人员针对主题演化做了大量研究,提出了众多主题演化可视化方法、工具。比如:S. Havre 等[9]提出 ThemeRiver 可视化模型,横轴表示时间,不同颜色的线条表示主题,并通过粗细表示主题在不同时间窗口下的强度,展示某领域的整体主题演化脉络;M. Rosvall 等[10]基于冲积图(Alluvial Diagram)提出一种社区主题演化可视化分析方法,将不同时期窗口下的社区展示在横向时间维度上,并以不同颜色的线条表示社区演化路径;王晓光等[11]开发了基于共词网络分析的主题演化可视化分析软件 Neviewer,提供赋色网络图、冲积图绘制功能,可以有效揭示主题演化的宏观过程和微观细节;牟冬梅等[12-13]将"三计学"理论、社会网络分析方法、学科知识结构理论和知识图谱技术进行集成优化和协同整合,并根据知识结构的高、中、低三个不同层级,针对性地探讨揭示不同层级知识结构的方法流程,为学科结构可视化分析提供了理论基础;同时牟冬梅等[14-15]利用时间-关键词共现分析构建时间-关键词二维矩阵,采用聚类分析、社会网络分析、时序词频统计和主题分类4种方法对时间-关键词二维矩阵进行可视化,对LIS领域学科动态知识结构进行多维度分析,并基于时序分析、主题-关键词共现分析构建2-模网络,利用 NetDraw 对各主题演化模式进行可视化呈现。

通过对现有的研究成果的分析发现:在主题识别方面,目前的研究大多进行静态主题识别,对于主题之间的相对重要性分析不足。实际上主题在不同时间段内主题之间存在主、次关系,将研究主题等同看待一定程度上限制了学科现状及其发展趋势分析的准确性和有效性;在主题演化方面,目前的研究侧重于通过分析主题强度、内容等不同维度的特征来分析其融合、分裂过程,但研究主题之间的关联关系识别以及主题关系在不同演化阶段的变化有待进一步深入研究;在主题演化路径可视化方面,目前的方法主要侧重对相邻时期主题的关联分析,对同一时间窗口下主题的相互关系的分析成果较少。针对目前研究中的局限,本文提出一种基于主题类别划分的主题识别及其演化路径可视化方法,对上述不足加以改善。

2 基于主题类别划分的研究主题识别及其演化路径可视化方法

2.1 方法的理论依据

期刊论文的关键词和主题词是其核心内容的提炼,研究主题是有效表征学科知识的基本单元。因此,可以通过文献计量、自然语言处理方法识别蕴含在期刊论文中的研究主题,分析某领域的热点、前沿和发展趋势。

美国海军研究所(Office of Naval Research, ONR)的 R. N. Kostoff 等[16]将研究主题分为普遍主题(pervasive themes)和副主题(sub-themes),通过实验分析了两种主题的关系:普遍主题和副主题具有紧密的关联关系,其中,当普遍主题发生变化会引起副主题的变化,但是副主题的变化基本不会引起普遍主题的变化;当普遍主题保持稳定时,副主题也可能会发生变化。普遍主题和副主题共同组成了完整的领域主题网络,在主题演化分析中,区分研究主题的主次关系,综合考虑两者的协同作用能够提升分析的准确性和有效性。

本研究借鉴 R. N. Kostoff 主题分析研究的基本思想,根据其提出的"普遍主题"和"副主题"概念,本研究中按照主题的重要程度将论文的主题划分为"核心主题"和"次要主题"两类,提出基于主题类别划分理论基础的核心主题识别及其演化路径可视化方法。

2.2 方法的流程与思路

基于主题类别的核心主题识别及其演化路径可视化方法基本步骤和思路为:第一步,根据领域确定数据源(数据库)、检索策略和时间区域等,进行期刊论文数据的收集整理;第二步,在数据预处理和划分时间窗口的基础上,利用LDA模型进行主题识别;第三步,结合多维尺度分析和可视化技术将LDA主题识别结果映射到二维空间,识别主题之间的关联关系,确定核心主题、次要主题;第四步,利用主题相似度算法,探测相

邻时期主题之间的关联,提出一种新的主题演化路径可视化方法,构建不同类型研究主题的交叉演化路径,以揭示核心主题、次要主题在演化过程中关系的动态变化。

下面对上述步骤中的主要内容进行具体介绍:

2.2.1 基于LDA模型的研究主题识别

近年来学界提出了不少主题模型,如潜在语义索引[17](Latent Semantic Analysis, LSA)、概率性潜在语义索引[18](probabilistic Latent Semantic Analysis, pLSA)和LDA模型等。与LSA和pLSA模型相比,LDA模型不仅能预测训练集文档的主题分布而且能够有效预测非训练集中的文档和词的主题分布,因此,LDA模型逐渐成为分析大规模非结构化文档集的主要的工具之一。

LDA是一种三层(词、主题和文档)贝叶斯概率模型,该模型假设文档是由若干隐性主题组成,而主题是由词表中的所有词汇组成。LDA主题模型的联合分布概率如公式(1)所示[19]:

$$P(\theta,z,w) - P(\theta|w)\prod_{n=1}^{N}P(z_n|\theta)P(w_n|z_n,\beta) \qquad (1)$$

其中,z表示主题,w表示主题词 N 表示第 m 个文档的单词数目,θ 为参数 α 的Dirichlet分布采样。由于LDA主题模型相较于其他主题识别方法(比如关键词聚类、社区探测等)能够更加准确、高效的分析文本主题,因此,本文之后将基于Python的scikit-learn工具包进行医疗健康信息领域的主题识别。

2.2.2 基于MDS的核心研究主题识别

LDA主题识别的结果一般难以直接分析不同主题之间的关联关系,为了获得研究主题中的核心主题,本文在上一步LDA主题识别结果的基础上,采用多维尺度分析(Multidimensional scaling, MDS),利用主题间的相似性构建低维空间,使得LDA主题在此空间的距离和在高维空间中的LDA主题之间的相似性尽可能的保持一致,从而可视化LDA主题的相互关系,直观地识别核心主题。

本研究中使用Python下的pyLDAvis工具包来绘制动态交互式的LDA主题可视化图谱,分析研究主题之间的关联关系,从而识别核心研究主题以及次要研究主题。pyLDAvis可以通过调节参数 λ($0 \leqslant \lambda \leqslant 1$)来控制主题－词语关联度relevance(term w | topic t),即可以控制显示某一主题的不同的下位词项。$\lambda=0$ 时,显示主题下特有的、相对独立的下位词项,即这些词项往往只出现在该主题;$\lambda=1$ 时,显示分布概率更高的下位词项,但是这些高分布概率的词项往往不单独属于该主题,也会同时属于其它主题。参数 λ 计算方法如公式(2)所示[20]:

$$r(w,k|\lambda) = \lambda\log(\varphi_{kw}) + (1-\lambda)\log\left(\frac{\varphi_{kw}}{p_w}\right) \qquad (2)$$

其中,w 表示主题词,$w \in \{1, 2,3...V\}$;k 表示主题,$k \in \{1, 2,3\cdots K\}$;φ_{kw} 表示Gibbs采样参数;p_w 表示主题词 w 的分布概率。

2.2.3 核心主题、次要主题演化路径可视化

在前文的主题演化路径可视化相关研究分析基础上可知,目前ThemeRiver、Textflow[21]和NEViewer等演化路径可视化方法模型主要侧重对相邻时期主题的关联分析,并且将所有主题等同对待,难以有效分析同一时间窗口下的主题的相互关系以及不同类型主题的演化关系。因此,本文提出一种新的领域核心研究主题识别及其演化路径可视化方法:基于R语言的流式图形分析核心主题、次要主题演化路径的可视化方法,该方法能够有效揭示核心主题、次要主题在演化过程中分裂、融合等关系的动态变化。

与现有的主题演化路径可视化方法模型相比,本文设计的演化路径可视化图谱能够分析某一类型研究主题随时间推移的流动模式,且可以分析核心主题、次要主题等不同类型研究主题之间的交叉演化脉络,展示关联关系的动态变化过程。

可视化的基本样式见图1,其中,块代表主题,块之间的流式图形代表随着时间的推移这些主题的演化路径(关联变化),粗细表示主题之间的关联强度;块的高度表示主题的强度(文献概率分布越高,主题块越大);核心研究主题添加"核心"标签,次要研究主题添加"次要"标签。

3 主题识别和主题演化可视化方法在医疗健康信息领域的应用

3.1 数据源及其预处理

本文选择CNKI期刊全文数据库作为数据源,收集题名、关键词和摘要等关键题录信息。具体检索策略如下:检索数据库: CNKI;检索策略:主题 ="医疗健康信息";时间跨度无限制;检索结果:704篇;检索时间:2018年6月3日。得到文献数量年度分布见图2。

目前研究者进行主题识别和演化分析,需要划分时间窗口以明确主题演化的时间维度(将期刊论文数据划分到若干连续的子时期),划分时间窗口的方法主

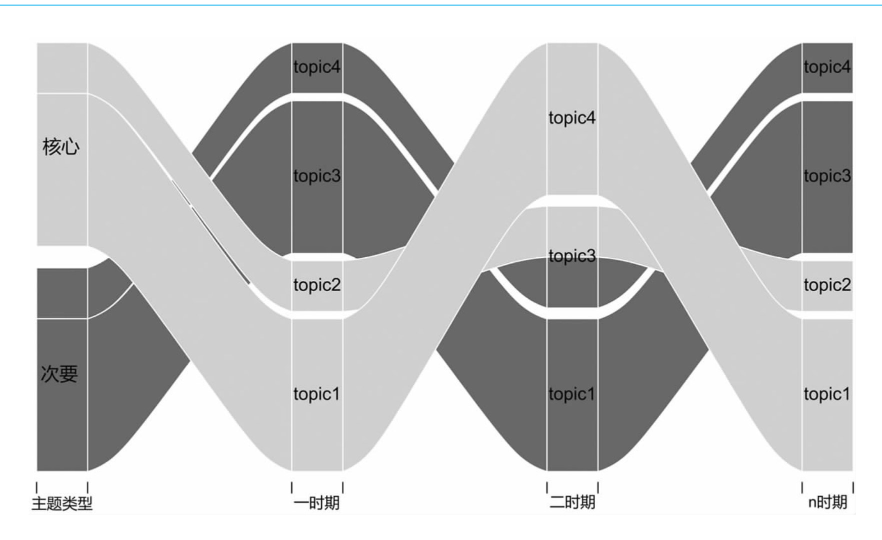


图 1　核心、次要主题演化路径可视化示例

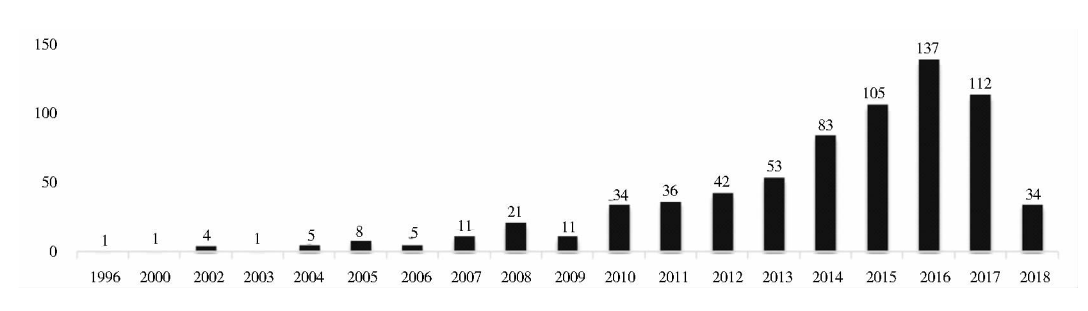


图 2　论文数量年度分布

要有根据数据的时间标签确定、按照年度固定时间窗口的两种方法[22-23]。本文采用按照年度固定时间窗口的方法，将检索结果划分为四个子时期，各个子时期年份和文献数量如表 1 所示：

表 1　主题识别与演化子时期划分

时期	年份(年)	文献数量(篇)
I	1996 – 2009	68
II	2010 – 2012	112
III	2013 – 2015	241
IV	2016 – 2018	283

由于 2010 年以前，国内医疗健康信息相关研究较少，因此划分 1996 – 2009 年为子时期 I(68 篇)，2010 年以后研究成果逐渐增多，所以每三年划分为一个子时期，2010 – 2012 年为子时期 II(112 篇)；2013 – 2015 年为子时期 III(241 篇)；2016 – 2018 年为子时期 IV(283 篇)。

由于主题识别仅需要分析文本字段，所以单独抽取出题名、关键词和摘要，对下载的四个子时期的期刊论文进行数据预处理，为主题识别做准备，处理内容主要包括格式转换、去重、去除停住词和标点符号等。

3.2　基于 LDA 模型的医疗健康信息研究主题识别

本文利用 Python 的 jieba 分词工具包进行中文分词后再进行 LDA 主题识别。为了提高分词结果的准确性，基于第一步收集的医疗健康信息领域期刊论文的关键词构建分词词典(关键词是期刊论文内容的高度概括与凝练，与 jieba 工具包自带的分词词典相比更加准确保持作者所想表达的主要词汇)，该词典的基本格式为词 – 词频 – 词性，由于本研究不涉及词性分析因此忽略词性，自定义词典以 txt 格式保存，通过 jieba.load_userdict(“dict. txt”)进行调用。

利用 Python 的 jieba 分词工具包进行中文分词后，利用 Python 的 scikit-learn 工具包进行 LDA 主题识别(按照所划分的四个时期依次进行 LDA 主题识别)。经过处理之后，各个子时期的 LDA 主题识别结果中，每个时期只展示了医疗健康信息领域相关的部分研究主题(本文只列举前 5 个，下位词表中展示部分)，后面是其对应的关键词和下位关键词，按照出现频率排序，见表 2。

3.3　基于 MDS 的医疗健康信息相关领域核心研究主题识别结果

为了更好地分析 LDA 主题之间的相互关系，在上一步 LDA 主题识别结果的基础上，基于多维尺度分析(Multidimensional scaling，MDS)构建 LDA 主题低维空间分布，来可视化 LDA 主题的相互关系，发现我国医

表 2　我国医疗健康信息相关领域不同时期研究主题及下位关键词列表(部分)

时期(年)	主题	关键词
Ⅰ时期(1996－2009)	健康档案	健康档案|健康传播|信息技术|区域医疗|老年人|贫困地区|大众传媒|发达国家|医疗服务
	健康平台	卫生局|B2C|健康网|医疗健康服务|看病难|统一标准|健康信息资源平台|安阳市|看病贵
	健康管理	健康管理|居民健康信息系统|消费者|医疗信息|特需医疗服务|非营利|调查报告|联合会
	健康信息工作	医疗健康信息|老年慢性病|健康信息|远程关怀|网络健康信息|成长之路|信息工作
	健康医疗信息	健康医疗信息|医疗档案|控制权|金卫网|隐私权|高速公路|医疗网络|个人信息|国家级|综合性
Ⅱ时期(2010－2012)	电子健康档案	电子健康档案|医疗机构|医疗服务|信息技术|云计算|健康信息|电子健康|健康档案
	互联网信息平台	一卡通|医疗保健|信息平台|医疗健康|中医临床信息标准|电子健康档案|体域网|互联网
	互联网健康传播	双向转诊|健康传播|健康信息|互联网|健康信息资源平台|健康监护|云平台|体域网
	健康生活方式	健康生活方式|宁波市|信息资源|物联网|医疗健康|医疗保健|电子健康|居民健康
	电子健康	电子健康档案|医疗机构|医疗服务|信息技术|云计算|健康信息|电子健康|健康档案
Ⅲ时期(2013－2015)	医疗健康信息化	医疗健康|健康档案|健康信息|信息化|移动健康|服务模式|云计算|大数据|居民健康档案
	电子医疗	老年人|医养一体化|信息平台|医疗信息|居家养老|医药电子商务|新媒体|医疗服务
	智慧医疗	HADOOP|智慧医疗|云计算|医疗服务|健康信息|个人健康信息管理|电子健康档案
	健康管理	健康管理|居民健康卡|云计算|健康信息|UGC|网络健康社区|互联网|社交媒体|健康教育
	移动医疗	移动医疗|APP|健康传播|健康信息|健康管理|医疗健康|移动互联网|穿戴式|信息平台
Ⅳ时期(2016－2018)	互联网医疗	医疗健康|互联网医疗|大数据|互联网|医疗服务|大数据应用|健康监测|医疗大数据
	电子健康素养	使用意愿|病患者|电子健康素养|智能健康管理|移动医疗服务|自我效能|O2O 医疗服务
	健康管理	健康管理|健康管理服务业|健康云|传染病患者|医疗服务|电子健康档案|服务平台
	健康信息获取	互联网|健康信息|健康信息获取|教职工|公共服务|信息技术|影响因素|医疗服务网站
	健康素养	健康信息|影响因素|健康素养|健康险|正确率|校医院|糖尿病|病患者|可穿戴设备

疗健康信息相关领域各个时期的核心研究主题。

具体数据处理过程是，基于 LDA 主题识别结果，使用 Python 下的 pyLDAvis 工具包分别绘制四个子时期的交互式 LDA 主题可视化图谱，如图 3、图 4、图 5 和图 6 所示（图中各时期主题列举 5 个，为手动添加；下位词列举 30 个）。

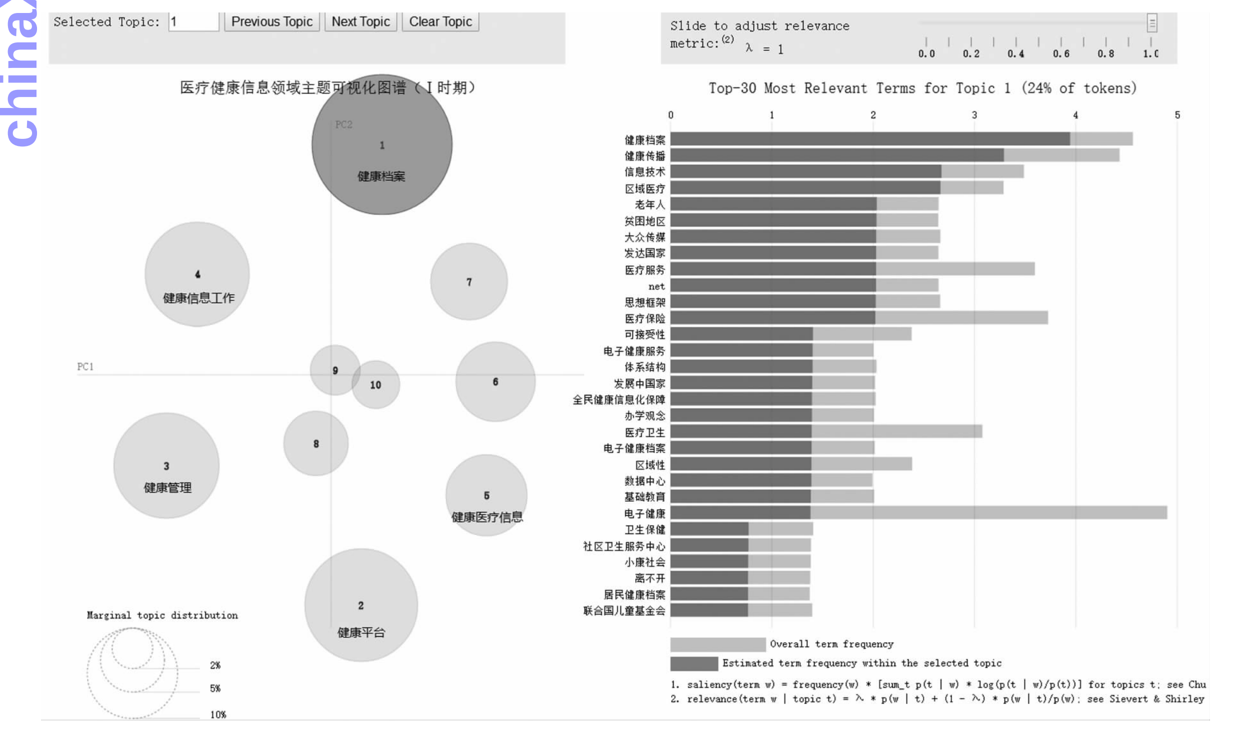


图 3　Ⅰ时期主题可视化

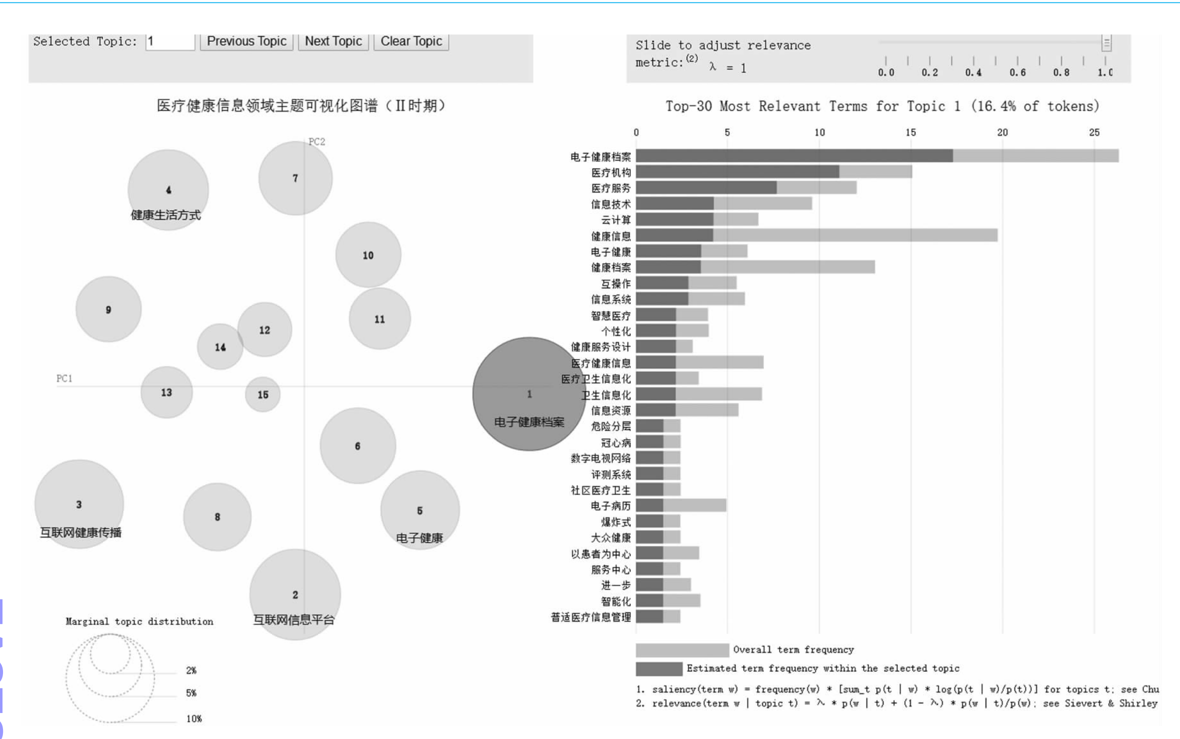


图 4　II 时期主题可视化

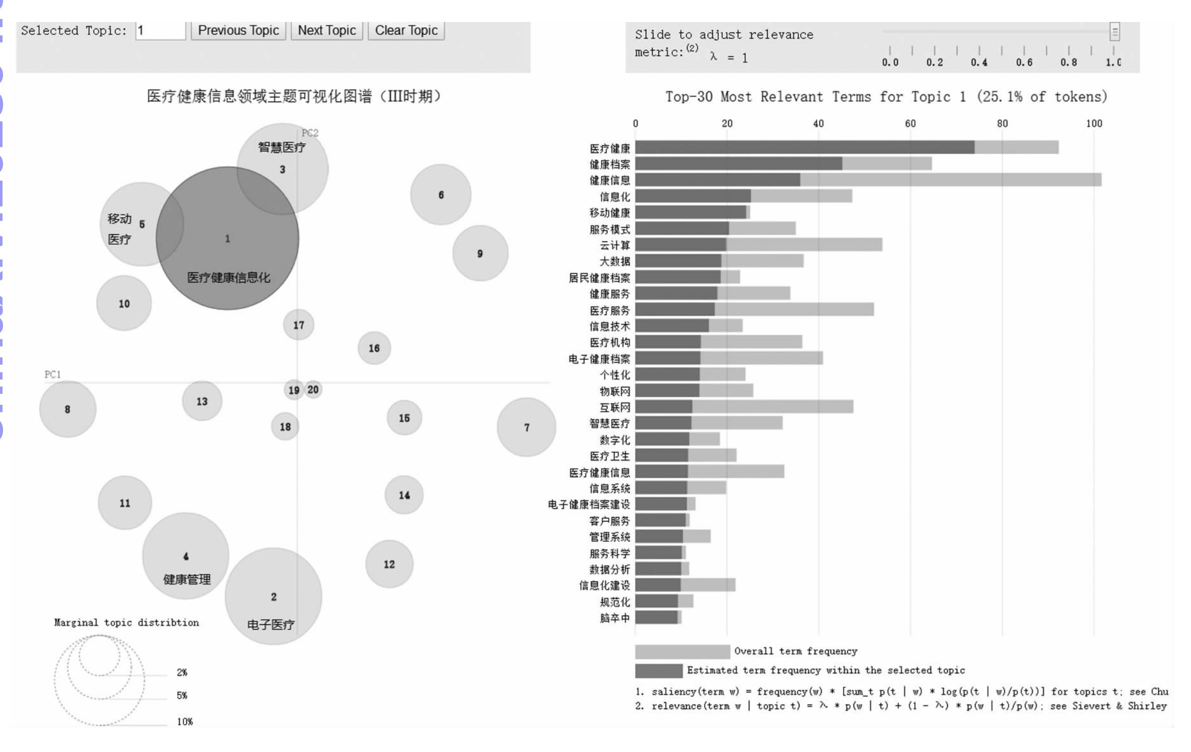


图 5　III 时期主题可视化

图 3、4、5、6 中左侧的大圈代表了核心主题，而小圈代表了次要主题，由于每个时期的文献和研究主题在数量上存在一定差距因此圈的数量也不同；右侧是每个主题的下位词项。可以据此对我国医疗健康信息领域不同时期的核心研究主题进行简要分析：

Ⅰ时期的核心主题为健康档案、健康平台、健康管理等，该时期医疗健康的相关研究逐渐开展，相关研究较为欠缺；Ⅱ时期的核心研究主题是电子健康档案、互联网信息平台、互联网健康传播等，由于技术的发展该阶段的研究主题基于新技术有了新的研究内容；Ⅲ时期的核心主题为医疗健康信息化、电子医疗、智慧医疗等，随着信息技术的进一步发展，电子健康进一步发展

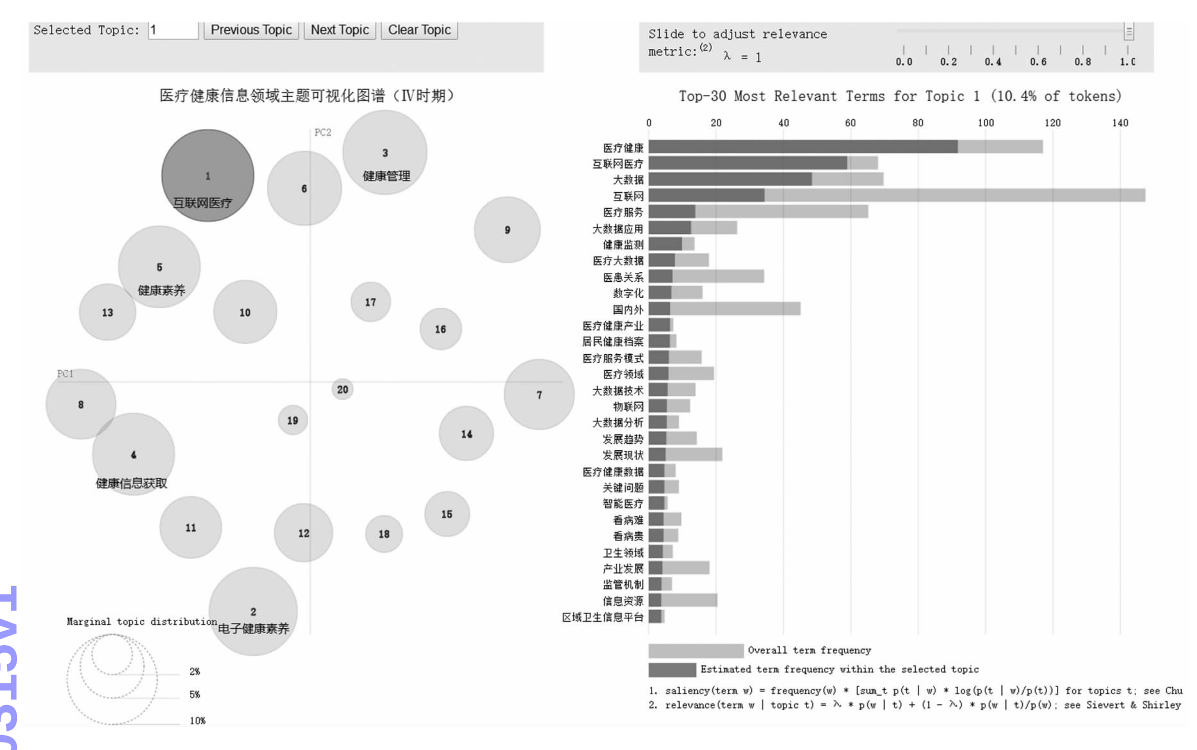


图6　Ⅳ时期主题可视化

并出现了智慧医疗；Ⅳ时期的核心研究主题是互联网医疗、电子健康素养、健康管理等，该时期依托技术，互联网医疗的研究进一步增多，并且公众对健康愈加重视，由于互联网技术的发展公众获取健康信息更加便利，因此关于健康素养的相关研究逐渐增多。

3.4　我国医疗健康信息相关领域研究主题演化可视化分析

3.4.1　核心主题、次要主题交叉演化分析

对我国医疗健康信息相关领域四个时期的核心研究主题识别结果基础上，基于本研究提出的核心主题、次要主题交叉演化路径可视化方法绘制演化脉络图，从而分析我国医疗健康信息领域在各个时期的核心研究主题及其发展演化过程，如图7所示（彩图网址：https://www.informationscience.top/topicevlution.html）。

图7中的核心主题以及次要主题选取的是基于LDA模型的医疗健康信息研究主题识别以及基于MDS的医疗健康信息核心研究主题识别的重叠主题（若某时期无此主题则计算主题相似度用相似度最高的相关主题代替），根据图7可以看出，每个时期的核心主题与次要主题是不断发展和变化的，下面将对几个代表性主题进行具体分析：

（1）电子健康档案。电子健康档案在Ⅰ、Ⅱ时期属于次要研究主题，随着技术的发展Ⅲ时期后转变为核心研究主题。目前该部分的研究主要包括两个方面，第一是国内外电子健康档案建设的对比研究以及国内的电子健康档案建设相关经验，介绍国外主要国家居民电子健康档案的共享服务体系建设，为我国共享服务体系建设提出建议；第二是电子健康档案和电子健康档案管理系统的建立，电子健康系统体系结构描述了电子健康档案的总体技术构成及其技术要素间的相互关系，是电子健康档案的核心技术之一。

（2）互联网医疗。互联网医疗是在IT技术的迅猛发展，移动通信进入4G时代，互联网应用演变至互联网+，大数据、云计算技术快速发展和普及的背景下发展起来的，因此与电子健康档案的演化脉络相似都是在Ⅲ时期后转变为核心研究主题。目前互联网医疗运用先进的信息化手段和互联网+应用平台提升医疗资源的使用效率，提高救治和服务水平成为近期我国医疗卫生行业发展的重要方向，新时期下技术因素在互联网医疗的应用推广过程中起着至关重要的作用，如何构建能长期健康持续发展的互联网医疗产业发展模式值得我们去研究。

（3）健康传播。健康传播除了在Ⅱ时期变为次要研究主题，其他时期皆为核心研究主题，该主题作为传播学的新兴分支，因其与个人生活的紧密关联和重大社会影响力而受到广泛关注，在每个时期都是比较重要的研究热点。但目前国内的健康传播研究[24]还停留在描述现象、个案讨论和概括此领域宏观特征的初

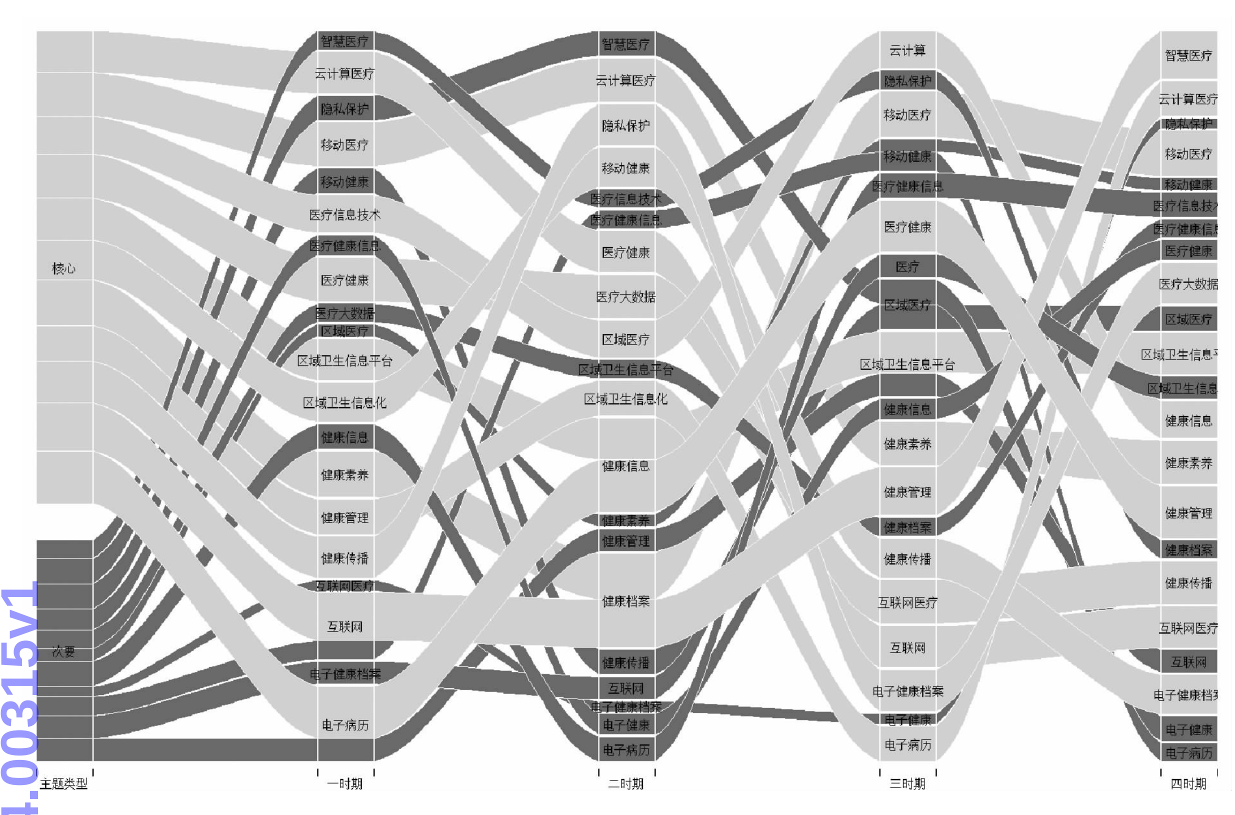


图 7　我国医疗健康信息相关领域核心、次要研究主题演化脉络

级阶段。探讨如何在理论关照下开展健康传播研究，探索健康信息传播过程中、健康行为发展过程中的重要影响因素；考察人与信息、人与人、人与社会的互动，以及健康信息传播带来的人的观念、态度、行为的变化将是十分重要的问题。

（4）移动医疗。移动医疗是指通过使用移动通信技术来提供医疗服务和信息，由于移动医疗的实用性较强，仅通过几个简单的 APP 就能实现，技术成本较低、简便易用，因而在演化阶段一直属于核心研究主题。目前研究主要集中于[25-26]：第一，移动医疗的具体实施过程研究进行详细的移动医疗健康需求分析总结出应用软件设计和移动医疗健康发展的重点；第二，梳理国内外移动医疗应用产业的现状，探究其典型应用的发展模式，并对关键要素展开分析，提出符合我国国情的发展建议；第三，对移动医疗的用户进行分析研究，探索面向不同用户的个性化移动医疗健康服务。

根据以上分析，研究主题类型的变化与技术的发展息息相关，电子健康档案、互联网医疗基于新技术逐渐成为医疗健康领域的核心研究主题；健康传播、移动医疗作为各个阶段的核心研究主题在技术的发展下有了新的研究内容；关于医疗健康信息的研究因为该主题与技术的相关度小于其他主题故而一直属于次要研究主题，但就目前的研究内容来看，该主题也逐渐开始利用新技术，在不久的将来也极有可能出现新的研究内容。

3.4.2　发展趋势

近年来，随着互联网技术的不断发展，医疗健康信息领域的研究主题也在不断变化，新技术主题不断涌现并呈现出不断增长的演化趋势，而部分主题由于新技术主题的冲击逐渐衰落，此外，部分重点研究主题依然保持良好的发展势头，在上一步核心、次要主题分析的基础上，对核心研究主题的发展趋势进行可视化分析，如图 8 所示（彩图网址：https://www.information-science.top/yh.html）。

根据图 8，选取 5 个典型的研究主题分析其发展趋势，这 5 个主题可以分为三类：第一类是医疗健康领域的重点研究主题，该类研究主题并未随着社会发展而消失，反而在新技术的影响下有了新的研究内容；第二类是技术发展背景下产生的新的研究主题，该类研究主题依赖于技术的发展，是时代发展的必然产物；第三类是消失的研究主题，该类研究主题的发展趋势也是受技术发展的影响，在新时期被更为先进的技术所取代，也是时代发展的必然过程。

（1）健康管理。健康管理一直以来都是医疗健康信息领域的研究重点，随着时间的演化呈现出逐渐增多的趋势。过去的期刊论文介绍国外健康管理的做法

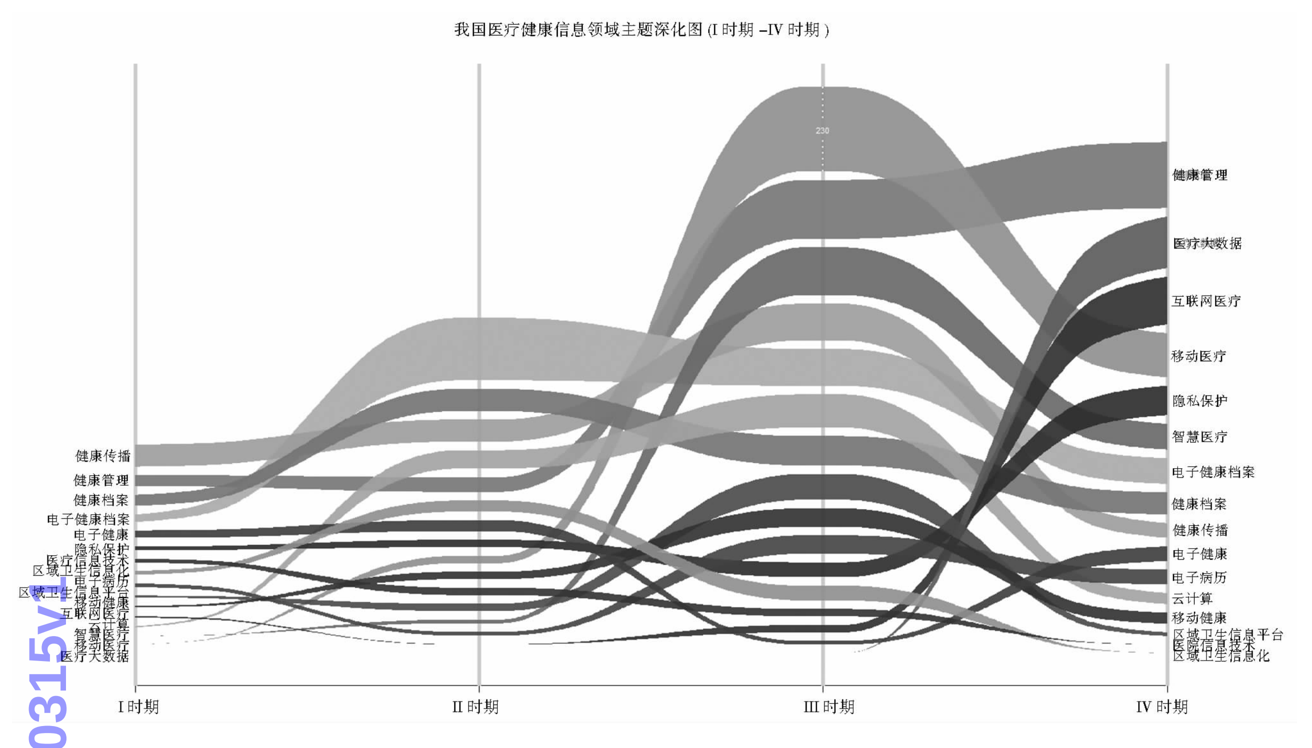


图 8　我国医疗健康信息相关领域核心研究主题发展趋势

和成绩比较多,随着近年来人们生活方式和健康理念的转变,期刊论文的研究成果体现了健康管理在新的健康模式下也表现出了新的特点,也逐渐成为一种新兴的健康服务理念和服务方式。健康管理一直是大众比较关注的领域,因此其在演化的过程中始终保持上升势头。健康管理服务业[27]以消费者健康需求为导向,以多元目标取代单一的经济目标,是人类自身进步与经济、社会协调发展的产业创新发展模式、人本发展模式,目前中国健康管理服务处于产业技术快速发展、多样化正在形成的初级阶段。在此背景下,健康管理服务的发展带动了学者对健康管理服务系统的研究,目前主要集中于以下几个方面:第一,以各类疾病的治疗为核心的医疗健康管理服务系统;第二,健康管理服务技术;第三,健康管理系统或体系的构成及比较研究。

(2)智慧医疗。智慧医疗目前主要有以下几个研究内容:第一,对我国智慧医疗的发展现状和趋势进行总结分析,并对国内外的建设现状进行对比,借鉴国外智慧医疗建设的有益经验提出相关可行的建设性建议;第二,介绍智慧医疗的起源和概念,从便捷的医疗服务体系、人性化健康管理体系、专业化的业务应用体系、科学化的监督管理体系、高效化的信息支撑体系、规范化的信息标准体系、常态化的信息安全体系几方面探索智慧医疗的应用,在此基础上提出发展建议。

智慧医疗是信息技术与医疗健康服务和管理的深入融合,并在国家的大力支持下取得了较快的发展,对医疗服务模式、卫生管理方式、居民健康管理等产生了深刻影响,但在医疗数据、系统安全、建设保障、资源共享、评价体系等方面仍存在一些问题与挑战。因此,智慧医疗在未来的发展道路上,仍需政府加强宏观指导、扩大信息共享范围,以更好地满足患者的需求。智慧医疗是近年来逐渐增多的研究主题,其出现和发展与互联网＋、大数据等技术密不可分,是新时代的产物。

(3)医疗大数据。医疗大数据是大数据环境下的必然产物,是近几年医疗健康信息领域新的研究热点与重点。医疗大数据作为大数据中极其重要的一部分,它的应用不仅仅是医疗方面的数据信息,还包括了卫生事业、生命健康数字化存储的海量数据。医疗大数据的研究主要为以下几个方面:第一,医疗大数据的研究现状及发展趋势的综述性研究。通过国内外医疗大数据的相关研究,明确国内医疗大数据的发展现状、研究热点并预测未来的发展趋势;明确医疗大数据在未来发展过程中可能遇到的机遇和挑战,制定相关措施实现最大化发展;医疗大数据对其他相关领域的贡献研究。第二,医疗大数据下的医疗服务模式。以建设临床数据中心为切入点,以临床数据中心为建设核心的医疗大数据平台,医疗大数据平台的建设研究成为新的热点[28]。第三,医疗大数据中的隐私保护问题

研究。因此医疗大数据的研究将是未来几年内的研究热点。

（4）医疗信息技术。医疗信息技术是将信息技术引入医疗领域，构建新的医疗服务模式，提供更好的医疗信息服务技术。医疗信息技术的研究最早出现于上个世纪九十年代，在该时期信息技术已逐渐开始向各个领域渗透。医疗领域顺应时代发展引入信息技术促进了医疗健康行业的发展。随着信息技术的进一步发展，医疗信息技术热点逐渐转向智慧医疗、医疗大数据等领域开展深度研究，期刊论文中单纯对医疗信息技术的研究逐渐减少。

（5）区域卫生信息化。区域卫生信息化与医疗信息技术的发展类似，技术的发展和公众的需求催生了该研究主题，同样因为技术的不断发展和公众日益迫切的需求而转向其他领域。区域卫生信息化的研究范围已经逐渐由小区域小范围上升到国家层面甚至国际层面，"区域"的含义已经逐渐发生改变，由更多的词汇代替"区域"一词，因此该研究主题的论文近年来出现了大幅缩减。

根据以上分析，二十年来医疗健康信息领域的研究主题发生了重大变化，研究主题的变化与技术发展有着密不可分的关系。新技术不断为医疗健康信息领域注入新的活力，对改善医疗服务模式、提高健康服务水平以及提升公众健康素养都有着重要的意义。但是技术的发展也在一定程度上带来了巨大挑战，公众隐私、医疗数据泄露等都极大地考验着从业者的专业素养，妥善利用新技术，为公众提供更好的服务应当是接下来重点思考的问题。

4 讨论与总结

本文中提出的方法借鉴 Kostoff 主题分析相关研究的基本思想，将研究主题划分为核心主题和次要研究主题，基于 MDS 构建 LDA 主题识别结果的关联关系探测核心研究主题，与目前基于 Citespace、UCINET 和 SPSS 等工具的核心研究主题识别及其可视化分析方法相比，本方法对研究主题之间的关联关系及其在不同演化阶段的变化作了进一步深入研究。此外，本文基于 R 语言提出一种针对核心主题、次要主题交叉演化的可视化方法，能够可视化展示领域研究主题的发展演化脉络，以及不同时间段内核心主题、次要主题的动态变化过程。基于大量科技文献数据的核心技术主题识别及其演化可视化方法，有助于识别某领域的核心研究内容、分析核心研究内容的发展方向，是进行科学创新的基础情报工作，具有较大的应用价值。

本研究主要存在两点局限：首先，LDA 主题识别结果（若干主题词的组合，解读困难）的解读依赖分析人员的专业知识，因此，需要探索更加有效的主题识别方法，提高结果的语义信息量，以便于解读；其次，研究中对于核心主题、次要主题的划分还有待进一步细化，比如次要研究主题可以分为新兴主题、衰退主题等。后续研究可以进一步探索利用语义增强的 LDA 模型进行主题识别以提高结果的可解读性，并尝试结合主题演化生命周期划分方法对主题类型进行多层次划分增加主题演化分析的维度。

参考文献：

[1] 王莉亚. 主题演化研究进展[J]. 情报探索, 2014(4):29-32.
[2] 白如江，冷伏海. k-clique 社区知识创新演化方法研究[J]. 图书情报工作，2013，57(17):94-99.
[3] RITZHAUPT A D. An investigation of distance education in North American research literature using co-word analysis[J]. International review of research in open & distance learning, 2010, 11(1):37-60.
[4] 程齐凯，王晓光. 一种基于共词网络社区的科研主题演化分析框架[J]. 图书情报工作，2013，57(8):91-96.
[5] 王效岳，刘自强，白如江，等. 基于基金项目数据的研究前沿主题探测方法[J]. 图书情报工作, 2017, 61(13):87-98.
[6] 李湘东，张娇，袁满. 基于 LDA 模型的科技期刊主题演化研究[J]. 情报杂志, 2014(7):115-121.
[7] 刘自强，王效岳，白如江. 多维度视角下学科主题演化可视化分析方法研究[J]. 中国图书馆学报，2016，42(6):67-84.
[8] 周源，张超，唐杰，等. 基于主题变迁的领域发展路径智能化识别[J]. 图书情报工作, 2018，62(14):62-71.
[9] HAVRE S, HETZLER E, WHITNEY P, et al. ThemeRiver: visualizing thematic changes in large document collections[J]. Visualization & computer graphics IEEE transactions on, 2002, 8(1):9-20.
[10] ROSVALL M，BERGSTROM C T. Mapping change in large networks[J]. PlosOne，2010，5(1):e8694.
[11] 王晓光，程齐凯. 基于 NEViewer 的学科主题演化可视化分析[J]. 情报学报，2013，32(9):900-911.
[12] 牟冬梅，郑晓月，琚沅红，等. 学科知识结构揭示模型构建[J]. 图书情报工作，2017，61(12):6-13.
[13] 郑晓月，牟冬梅，琚沅红，等. 学科知识结构揭示流程与方法探究[J]. 图书情报工作，2017，61(12):14-20.
[14] 牟冬梅，琚沅红，郑晓月，等. 基于时间-关键词共现分析的学科动态知识结构研究——以国外图书情报学为例[J]. 图书情报工作，2017，61(12):21-31.
[15] 郑晓月，牟冬梅，琚沅红，等. 学科知识结构主题演化模式研究——以图书情报学领域"计量学"主题为例[J]. 图书情报工作，2017，61(12):32-41.

[16] KOSTOFF R N,EBERHART H J,TOOTHMAN D R,et al. Database tomography for technical intelligence:comparative roadmaps of the research impact assessment literature and the Journal of the American Chemical Society[J]. Scientometrics,1997,40(1):103－138.

[17] LANDAUER T K, DUMAIS S T. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge[J]. Psychological review, 1997, 104(2):211－240.

[18] SHEN C, LI T, DING C H Q. Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (PLSA) with sentence bases[C]// AAAI Conference on Artificial Intelligence. Palo Alto:AAAI Press, 2011:914－920.

[19] BLEI D M,Ng A Y,JORDAN M I. Latent Dirichlet allocation[J]. The journal of machine learning research,2003(3):993－1022.

[20] SIEVERT C, SHIRLEY K. LDAvis: a method for visualizing and interpreting topics[C]// Proceedings of the workshop on interactive language learning, visualization, and interfaces. Baltimore: Association for Computational Linguistics, 2014:63－70.

[21] CUI W W,LIU S X,LI T,et al. Text flow:towards better understanding of evolving topics in text[J]. Transactions on visualization and computer graphics,2011,17(12): 2412－2421.

[22] 吴斌,王柏,杨胜畸. 基于事件的社会网络演化分析框架[J]. 软件学报, 2011,22(7):1488－1502.

[23] 钱铁云,李青,许承瑜. 面向科技主题发展分段的社区核心圈技术[J]. 计算机科学与探索, 2010,4(2):170－179.

[24] 赵加奎,林军,陆瑛,等. 新媒体在健康传播中的应用现状及对策[J]. 中国健康教育,2016,32(10):919－921.

[25] 胡蓉,陈惠芳,徐卫国. 移动医疗服务中医患互动对患者感知价值的影响——以知识共享为中介变量[J]. 管理科学,2018,31(3):75－85.

[26] 王雪,石元博,黄越洋. 基于 Apriori 算法的数据挖掘在移动医疗终端系统中的研究[J]. 数字技术与应用,2017(9):60－61.

[27] 韩正臣,张基栋. 基于安卓的诊所预约与健康管理系统的设计与实现[J]. 信息技术与信息化,2016(12):30－35.

[28] 林静,吴向阳. 基于"互联网＋"的医疗信息化建设[J]. 电脑知识与技术,2016,12(23):216－218.

作者贡献说明：
岳丽欣：负责主题识别及可视化相关文献调研与资料收集，论文撰写与修改；
周晓英：负责论文框架和思路设计，指导论文撰写并修改论文；
陈旖旎：论文修改。

Research on Topic Identification of Papers Core Research Subjects and Evolution Path Visualization Method
——Taking China's Journal of Medical and Health Information as an Example

Yue Lixin　Zhou Xiaoying　Chen Yini
School of Information Resources Management, Renmin University of China, Beijing 100872

Abstract: [**Purpose/significance**] This paper proposes the identification of the core research topics and their evolution path visualization methods, in order to provide reference for the field subject evolution analysis research, which has certain significance for revealing the evolution characteristics and development laws of the core topics. [**Method/process**] Using the LDA model for topic recognition and combining multi-dimensional scaling analysis and visualization techniques to map LDA topic recognition results to two-dimensional space. The topic similarity algorithm was used to detect the association between adjacent time topics, a new visual display method was proposed. We constructed cross-evolution paths of different types of research topics to reveal the dynamic changes of core topics and secondary topics in the evolution process. [**Result/conclusion**] Taking the medical health information field in China as an example, the research results show that the core research topics in the field of medical and health information in China mainly include electronic health records and Internet medical treatment. Among them, core themes such as health management and smart medical treatment show a good development trend.

Keywords: core research topics　topic recognition method　topic evolution path　visualization　medical health information